DXE_EMTR 2022
First assignment (20% of total grade)

Please submit the assignment by 4.11.2022 in the IS MUNI system. You are allowed to work in groups of maximum size 3.

For students who were not exposed to programming before or who wish to increase their familiarity with software `R`, attending `DXE_EREC` course is highly recommended.

1 Regression basics (A note on screening regression equations)

We have the following setup: 1000 observations, outcome y and 200 regressors $x_1, x_2, ..., x_{200}$. Our aim is to explore if the variation in $x_1, ..., x_{200}$ can explain variation in y.

Consider the following algorithm:

Step 1 Run a linear regression explaining y using all 200 regressors (include a constant term).

Step 2 Rerun the linear regression (include a constant term) but only keep the regressors with p-value smaller than 0.5 in Step 1.

Step 3 Rerun the linear regression (include a constant term) but only keep the regressors with p-value smaller than 0.25 in Step 2.

Step 4 Rerun the linear regression (include a constant term) but only keep the regressors with p-value smaller than 0.1 in Step 3.

Step 5 Calculate how many regressors are significant at level 0.05 in Step 4.

Now generate 201000 independent normally distributed $N(0, 1)$ random draws. These numbers will constitute vectors y and $x_1, x_2, ..., x_{200}$. Note that by design, there is no relationship between y and any of these $x_1, ..., x_{200}$. Apply the above algorithm to this data.

- How many regressors are significant in at 0.05 in the last step?
- Repeat this many process many times (e.g. 1000) and plot the distributions of the number of significant regressors in Step 5.
- Explain your findings.
- Discuss what does this imply for a statistical practice, in other words what is the lesson to take.
- (*) Compare the distributions of the numbers of significant regressors in the different Steps of the algorithm. Explain why do they differ.

Potentially useful reading: Freedman, David A. 'A note on screening regression equations.' The American Statistician 37.2 (1983): 152-155.

2 Maximum likelihood

There are situations in which our dependent variables describe a number of successful independent trials, out of some value n_i - the number of total trials. In this case $Y_i \in \{0, 1, 2, ..., n_i\}$. This may be, for instance, a number of failed o-rings in the Challenger example from the lecture, a number of Titanic survivors, a number of successful start-ups.. Our ambition may be to find an association between the variation in the y and some explanatory variables $x_1, \cdots x_p$, these may include weather conditions, geographic location or market conditions, depending on what the variable y is.

Consider the special type of regression (called Binomial regression), where we assume the following assumptions:

- Data sample consists of n i.i.d. observations $(y_i, n_i, x_{i1}, \cdots, x_{ip})$,
- $y_i \sim Bin(n_i, p_i)$,
- $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}$,

You are asked to do the following:

- We are interested in estimating the vector of unknown parameters $(\beta_0, \beta_1, \cdots, \beta_p)$. Derive the likelihood function (conditional on the covariates $x_1, \cdots, x_p$) and the score function for this model.
- Using a small simulation study in `R`:
  - demonstrate that the maximum likelihood estimator of β_1 for this particular model has asymptotically normal distribution. (You don't need to implement the optimization yourself, it is OK if you make use of `glm` function in `R` with option `family = binomial`, please check the documentation for the correct syntax).
  - Explore how the sample size affects the variance of the estimator. Visualize this relationship and compare it with theoretical predictions.

3 Bootstrap

Consider the maximum likelihood estimator of the unknown parameter β_1 from the previous task.

- Construct a 95% confidence interval based on the non-parametric percentile bootstrap.
- Using a simulation study in `R`, compare the coverage properties of these two confidence intervals. That is: show that the confidence intervals cover the true value in approximately 95% simulated cases.[1]

Submit your own work. Make sure that your code runs without errors. Make sure to comment your code and make your best effort to adhere to some reasonable coding standards. Your code must be easy to read. Present your results in a coherent way and whenever possible make use of visualization.

[1] Notice that this may require a lot of computing time - if a single bootstrap confidence interval is based on 100 bootstrap samples and you will run 500 simulations, you will need to estimate the Poisson regression model $100 \cdot 500 = 50000$ times. So you may need to keep the basic model specification simple.